JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 16, Number 6, 2009
© Mary Ann Liebert, Inc.
Pp. 803–815
DOI: 10.1089/cmb.2008.0219

An $O(n^5)$ Algorithm for MFE Prediction of Kissing Hairpins and 4-Chains in Nucleic Acids

HO-LIN CHEN,[1] ANNE CONDON,[2] and HOSNA JABBARI[2]

ABSTRACT

Efficient methods for prediction of minimum free energy (MFE) nucleic secondary structures are widely used, both to better understand structure and function of biological RNAs and to design novel nano-structures. Here, we present a new algorithm for MFE secondary structure prediction, which significantly expands the class of structures that can be handled in $O(n^5)$ time. Our algorithm can handle H-type pseudoknotted structures, kissing hairpins, and chains of four overlapping stems, as well as nested substructures of these types.

Key words: kissing hairpins, pseudoknot, RNA, secondary structure prediction.

1. INTRODUCTION

Our knowledge of the amazing variety of functions played by RNA molecules in the cell continues to expand, with the functions determined in part by structure (Lee et al., 1997). Additionally, DNA and RNA sequences are designed to form novel structures for a wide range of applications, such as algorithmic DNA self-assembly (He et al., 2008; Rothemund et al., 2004), detection of low concentrations of other molecules of interest (Dirks and Pierce, 2004) or to exhibit motion (Simmel and Dittmer, 2005). In order to improve our ability to determine function from DNA or RNA sequences, and also to aid in the design of nucleic acids with novel structural or functional properties, accurate and efficient methods for predicting nucleic acid structure are very valuable.

Currently, computational prediction methods focus mostly on secondary structure—the set of base pairs that form when the molecule folds. Of particular interest, from a computational standpoint, is prediction of pseudoknotted secondary structures—those in which base pairs *cross*, as illustrated in Figure 1. Biologically important examples of pseudoknots include H-type pseudoknots (ABAB motif), kissing hairpins (ABACBC motif), and such structures with nested substructures (Fig. 1).

A common approach to prediction of nucleic acid secondary structure from the base sequence is to find that structure with the minimum free energy (MFE), from the possibly exponentially many secondary structures that the molecule can form (Akutsu, 2000; Dirks and Pierce, 2003; Mathews et al., 1999; Reeder and Giegerich, 2004; Rivas and Eddy, 1999; Uemura et al., 1999). The energy of a structure is modeled as

[1]Department of Electrical Engineering, California Institute of Technology, Pasadena, California.
[2]Department of Computer Sciences, University of British Columbia, Vancouver, British Columbia, Canada.

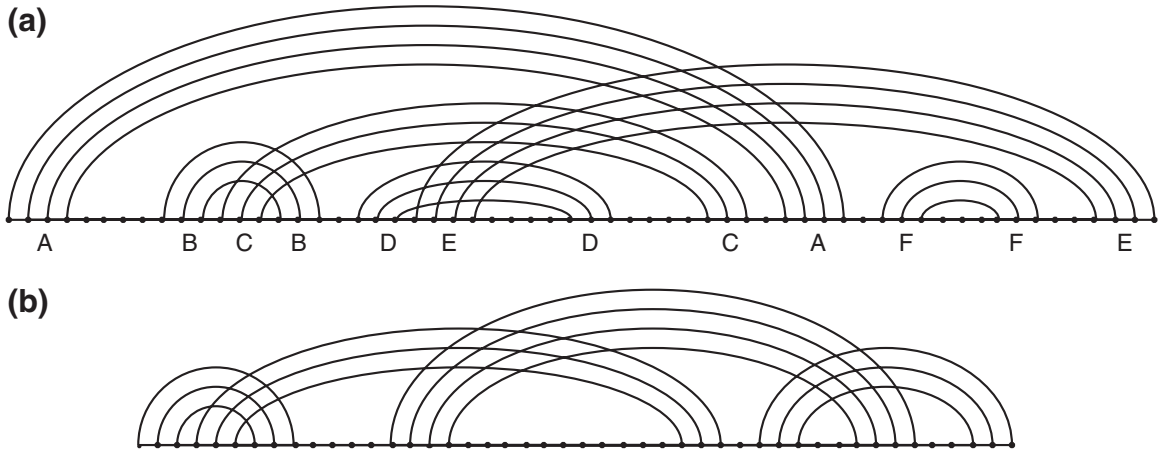

FIG. 1. (**a**) Arc diagram representation of a simplified version of the structure the aptamer core of a SAM-IV riboswitch (Weinberg et al., 2008), but preserving patterns of crossing base pairs. Dots along the horizontal line represent bases from the 5′ (left) to the 3′ (right) ends. Arcs represent base pairs. Some arcs cross, thereby introducing pseudoknots. Below both the left and right endpoints of each stem is a letter which identifies that stem; concatenating the letters from left to right yields the motif for this structure, namely ABCBDEDCAFFE. (**b**) A chain of four overlapping stems. This structure has motif ABACBDCD. If either the leftmost or rightmost stem were removed from this structure, the result would be a kissing hairpin structure, with motif ABACBC.

the sum of the energies of loops formed by the base pairs, and the loop energies are provided by a table or by a formula that takes into account the number of unpaired bases or base pairs around the loop. Given an MFE algorithm for structure prediction, it is often possible to adapt the algorithm to calculate a quantity called the partition function, which in turn leads to methods for calculating probabilities of base pair formation. These probabilities make it possible to associate confidences with particular base pair predictions or to identify more than one stable secondary structure for a sequence (Dirks and Pierce, 2003; McCaskill, 1990).

Since the general problem of MFE nucleic acid secondary structure prediction is NP-hard (Akutsu, 2000; Lyngsø, 2004; Lyngsø and Pedersen, 2000), polynomial-time algorithms handle—that is, find the MFE structure from—a restricted class of secondary structures. For example, the MFE pseudoknot free secondary structure for a sequence can be found in $O(n^3)$ time, where n is the length of the molecule. A very general algorithm, with running time of $\Theta(n^6)$ and space $\Theta(n^4)$, was proposed by Rivas and Eddy (1999), which handles most known biologically important structures, including those of Figure 1.

There are also several algorithms for predicting MFE pseudoknotted secondary structures that run in $\Theta(n^5)$ time and $\Theta(n^4)$ space (Lyngsø and Pedersen, 2000; Uemura et al., 1999). All of these algorithms handle a more limited class than does the Rivas and Eddy algorithm. All can handle H-type pseudoknots (ABAB motif), and some can handle kissing hairpin structures (ABACBC motif) when these do not have arbitrary nested substructures (Uemura et al., 1999). There are also some algorithms that run in $\Theta(n^4)$ time (Lyngsø and Pedersen, 2000; Reeder and Giegerich, 2004); these handle classes of structures that are even more restricted than the $\Theta(n^5)$ algorithms. However, none of the algorithms to date handle kissing hairpin structures with arbitrary nested substructures, and this is a real limitation given the biological importance of such structures. For example, neither of the structures of Figure 1 can be handled by any of the algorithms with $O(n^5)$ running time.

In this article, we present new algorithmic ideas that significantly expand the class of structures that can be predicted in $O(n^5)$ time and $O(n^4)$ space. Example of motifs that our algorithm can handle, but cannot be handled by previous $O(n^5)$ algorithms, include the structures shown in Figure 1, namely ABCBDEDCAFFE and ABACBDCD, as well as ABADDCBEEC. The algorithm can also handle arbitrary nested substructures of these types, as well as nested kissing hairpins.

The rest of our article is organized as follows. In Section 2, we review background on secondary structures and present our energy model. We also describe the types of structures that our methods can handle. We present our algorithm in Section 3, along with an illustrative example. We conclude with a summary and suggestions for future directions in Section 4.

2. BACKGROUND ON NUCLEIC ACID SECONDARY STRUCTURE

Here, we provide a short overview of nucleic acid secondary structures and their components.

An RNA sequence is a string over {A,C,G,U}. The pairs of bases (A,U), (C,G), (G,C), (G,U), (U,A), and (U,G) are said to be *complementary* base pairs. A DNA sequence is defined similarly, with T replacing U. In what follows, let S be an RNA or DNA sequence of length n.

A *base pair* for S is an ordered pair $i.l$ with $1 \le i < l \le n$, such that ith and lth bases of S are complementary. We call i and l the *endpoints* of the base pair. The *span* of base pair $i.l$ is $l - i$. If $i.l$ and $d.e$ are two base pairs with $i \le d$, then (i) $i < d < e < l$, in which case $d.e$ is *nested* in $i.l$; (ii) $i < l < d < e$, in which case $i.l$ and $d.e$ are *disjoint*, (iii) $i < d < l < e$, in which the base pairs *cross*, or (iv) either $i = d$, $d = l$, or $l = e$, in which case the base pairs *collide*.

A *secondary structure* R for S is a set of base pairs for S, none of which collide. With respect to R, base pair $i.l$ of R is *pseudoknotted* if it crosses some base pair of R. R is *pseudoknotted* if it contains pseudoknotted base pairs, and is pseudoknot free otherwise.

When $i \le l$, a *region* $[i, l]$ is the set of indices between i and l, inclusive and when $i > l$, the region $[i, l]$ is the empty set. A *gapped region* is the union of two regions $[i, j]$ and $[k, l]$ with $i < j + 1 < k \le l$. Base pair $d.e$ *spans* gapped region $[i, j] \cup [k, l]$ if $d \in [i, j]$ and $e \in [k, l]$. With respect to secondary structure R, we say that region $[i, l]$ is *structure-free* if no base in the region is paired in R. Region $[i, l]$ is *weakly closed* if no base pair connects a base in the region to a base outside the region. Region $[i, l]$ is *a pseudoknot* if it is weakly closed, both i and l are paired but not to each other, and the region cannot be partitioned into distinct weakly closed regions.

Base pairs of secondary structure R partition the unpaired bases of sequence S into *loops* (Rastegari and Condon, 2007). Loops, and their associated unpaired bases and closing base pairs, are defined as follows.

- A *hairpin loop* has one closing base pair $i.l$; all bases in $[i, l]$ are unpaired and belong to the loop.
- An *internal loop* has an external closing base pair $i.l$ and a second closing base pair $d.e$ which is nested in $i.l$. All bases in regions $[i + 1, d - 1]$ and $[e + 1, l - 1]$ are unpaired and belong to the loop. We say an internal loop has an asymmetry of $z > 0$ if the absolute value of the difference between the number of unpaired bases on one side of the loop and on the other side of the loop is z. A special case is a *stacked pair*, in which $d = i + 1$ and $e = l - 1$.
- A *multiloop* with external base pair $i.l$ arises in two cases. In the first, some base pair $d.e$ is nested in $i.l$; also regions $[i + 1, d - 1]$ and $[e + 1, l - 1]$ are weakly closed but are not both structure-free. In the second, $[d, e]$ is a pseudoknot and regions $[i + 1, d - 1]$ and $[e + 1, l - 1]$ are weakly closed (and may both be structure-free).
- An unpaired base u in $[i, l]$ belongs to the multiloop if u *can see* the base pair $i.l$ — that is, there is no base pair $x.y$ in R with $i < x < u < y < l$. The closing base pairs of the multiloop are $i.l$, the base pair $d.e$ if in the first case, plus any non-pseudoknotted base pair $x.y$ in $[i + 1, d - 1] \cup [e + 1, l - 1]$, where both x and y can see $i.l$.
- A *pseudoloop* is associated with each pseudoknot $[i, l]$. Its closing base pairs are of two types. Base pairs of the first type are the borders of the pseudoknot's *bands*. Here, a *band* is a maximal set of pseudoknotted base pairs, all of which cross exactly the same set of base pairs, and its *borders* are the base pair(s) with the largest and smallest spans. We refer to the border with the largest span as the external border. A band *belongs to* pseudoknot $[i, l]$ if its base pairs are within $[i, l]$ and are not within any pseudoknot nested within $[i, l]$. Closing base pairs of the second type are any unpseudoknotted base pairs whose endpoints can see at least two bands of the pseudoknot. Finally, unpaired bases of the pseudoloop are those unpaired bases in $[i, l]$ that do not belong to any other loop within $[i, l]$.
- An *external loop* contains all of those unpaired bases in $[1, n]$ that are not in any other loop.

A *stem* is a maximal sequence $bp_1, bp_2, \ldots, bp_k$ of base pairs, where successive base pairs bp_i and bp_{i+1} are the closing base pairs of an internal loop. Note that a band may be composed of several stems, separated by multiloops. We can obtain a *motif* for a secondary structure, which describes the pattern of overlaps of its stems, in the following way. Label each stem with a distinct symbol, write each symbol under both the left and right ends of the stem in the arc diagram for the structure (so each symbol appears twice), and concatenate the symbols in order from left to right. A structure is an H-type pseudoknot if its motif is ABAB, a kissing hairpin structure if its motif is ABACBC, and a chain of four overlapping stems if its motif is ABACBDCD (some of these types are illustrated in Fig. 1). If the motif for a structure has substring AxByAzB, where A and B are symbols and x, y, and z are arbitrary strings with no symbol in common, not all of which are empty, then we say that the structure has an H-type pseudoknot with nested substructures. Similarly, a structure contains a kissing hairpin with nested substructures if its motif has a substring of the form AvBwAxCyBzC, where A, B and C are symbols and v, w, x, y, and z are arbitrary strings with no symbol in common, not all of which are empty.

TABLE 1. ENERGY PARAMETERS AND FUNCTIONS

Name	*Description*
$e_H(i,l)$	energy of a hairpin loop closed by $i.l$
$e^p_{stacked}(i,l)$	energy of a stacked pair $i.l,(i+1).(l-1)$ that spans a band
$e_{int}(i,d,e,l)$	energy of an ordinary internal loop closed by $i.l$ and $d.e$
α^p_0	penalty for initiation of internal loop that spans a band
$\alpha^p_1(z)$	penalty for z unpaired bases in an internal loop that spans a band
$\alpha^p_2(i,l)$	penalty for closing pair $i.l$ or $l.i$ of an internal loop that spans a band
$\alpha^p_3(z)$	penalty for asymmetry of z in an internal loop that spans a band
β_0	penalty for initiation of an ordinary multiloop
β^p_0	penalty for initiation of multiloop that spans a band
β_1	penalty for unpaired base of an ordinary multiloop
β^p_1	penalty for unpaired base of a multiloop that spans a band
$\beta_2(i,l)$	penalty for closing pair $i.l$ or $l.i$ of an ordinary multiloop
$\beta^p_2(i,l)$	penalty for closing pair $i.l$ or $l.i$ of a multiloop that spans a band
γ_0	penalty for initiation of an external pseudoloop
γ^m_0	penalty for initiation of pseudoloop in a multiloop
γ^p_0	penalty for initiation of pseudoloop in a pseudoloop
γ_1	penalty for unpaired base of a pseudoloop
$\gamma_2(i,l)$	penalty for closing pair $i.l$ or $l.i$ of a pseudoloop

The values of functions such as $\beta_2(i,l)$ typically depend on the bases in positions $i, l, i+1$ and $l-1$ of S, and whether $i<l$.

2.1. Energy model

We use a loop-based energy model (Mathews et al., 1999; Zuker and Sankoff, 1984; Zuker and Stiegler, 1981). This model is used in many software packages, such as Mfold (Zuker, 2003) and the Vienna RNA package (Hofacker et al., 1994). In this energy model, the energy of a secondary structure is the sum of the energy of the structure's loops. Several parameters and functions inform the energy model; we summarize these in Table 1 and assume that any function specified there can be calculated in constant time, given sequence S.

The energy associated with hairpin loop closed by $i.l$ is denoted by $e_H(i.l)$.

The energy of an internal loop or multiloop depends on whether or not the external base pair of the loop is pseudoknotted. If it is not, we call the loop *ordinary*, and otherwise say that the loop *spans a band*. The energy of an ordinary internal loop closed by $i.l$ and $d.e$ is denoted by $e_{int}(i,l,d,e)$. We need to be more restrictive in the form of the energy for an internal loop that spans a band. If the loop is a stacked pair, its energy is given by $e^p_{stacked}(i,l)$. Otherwise, if the external base pair is $i.l$ and the other closing base pair is $d.e$, then the number of unpaired bases in the loop is $U=(d-i-1)+(l-e-1)$, and the asymmetry is $A=|(d-i-1)-(l-e-1)|$. The energy is then

$$\alpha^p_0+\alpha^p_1 U+\alpha^p_2(l,i)+\alpha^p_2(d,e)+\alpha^p_3(A).$$

Note that, here, the order of the parameters to the function α_2 differs, depending on whether or not the parameters identify the external base pair. That is, for the external base pair $i.l$, the larger index, l, is the first parameter to α^p_2 whereas for the other closing base pair $d.e$, the smaller index, d, is the first parameter to α^p_2. This is because in some energy models, the penalty for the closing base pair of an internal loop that spans a band may also depend on bases within the loop that are adjacent to the closing base pair. If the closing base pair is the external base pair $i.l$, then the adjacent bases are $i+1$ and $l-1$, and if the closing base pair $d.e$ is not the external base pair, then the adjacent bases are $d-1$ and $e+1$. In either case, the order of the parameters to $\alpha^p_2(x,y)$ ensures that the base adjacent to the first parameter, x, is $x-1$ and the base adjacent to the second parameter, y, is $y+1$.

The energy associated with an ordinary multiloop which has U unpaired bases, external base pair $i.l$, and set of other closing base pairs C is

$$\beta_0+\beta_1 U+\beta_2(l,i)+\sum_{d.e\in C}\beta_2(d,e).$$

As was the case for internal loops, if $\beta_2(x, y)$ is a term in this formula, then either $x.y$ or $y.x$ is a base pair in the loop and the bases within the loop which are adjacent to this base pair are $x-1$ and $y+1$. The energy associated with a multiloop which spans a band is similar, with β_x^p replacing $\beta_x, 0 \le x \le 2$.

As shown next, the energy of a pseudoloop with U unpaired bases and set of closing base pairs C is also similar, with the γ parameters replacing the β parameters. Just as the previous formulas distinguish between external closing pairs and other closing pairs, in the next formulas we distinguish between external band borders and other closing pairs. Let C_E be the set of closing pairs of a pseudoloop which are external band borders, and let C be the set of all other closing pairs of a pseudoloop, plus the band borders of bands which have only one base pair. For an external pseudoloop—that is, a pseudoloop which is not nested in any other type of loop—the energy is

$$\gamma_0 + \gamma_1 U + \sum_{i.l \in C_E} \gamma_2(l, i) + \sum_{i.l \in C} \gamma_2(i, l).$$

If the pseudoloop is within a multiloop, γ_0 is replaced by γ_0^m, and if it is within another pseudoloop, γ_0 is replaced by γ_0^p. Finally, the energy associated with an external loop is 0.

2.2. *The CCJ class of structures*

We call the class of structures that can be handled by our algorithm *CCJ structures* (using the initials of the last names of the co-authors of this paper). To explain what CCJ structures are, we also introduce TGB (three-groups-of-bands) structures.

Figure 2 illustrates these structures. Part (a) shows a TGB structure for a gapped region, which is comprised of three groups of bands. Bands in the middle group cross bands in the left and right groups according to a regular pattern. More generally, a TGB structure always has at most three groups of bands. Also, a TGB structure always has at least one band in the middle group which spans the gap, and for each base pair $i.l$ in this middle band, the base at position i is always nested in every left band of the group (if any), and the base at position l is always nested in every right band of the group (if any). A TGB structure may have nested substructures, either within a band (as illustrated in part (b)), or between bands. A CCJ pseudoknot is an "overlay" of two disjoint TGB structures, as shown in part (c). An H-type pseudoknot is a CCJ pseudoknot, as is a kissing and even a chain of four overlapping stems. Finally, CCJ structures can be pseudoknot free, can have CCJ pseudoknots, and can have nested CCJ structures interspersed in arbitrary places.

In the rest of this section, we define CCJ and TGB structures inductively. If R is a secondary structure for a sequence of length n, let $R_{[i,l]}$ be the subset of base pairs of R whose endpoints are in region $[i, l]$, and let $R_{[i,j]\cup[k,l]}$ be the subset of R whose endpoints are in $[i,j] \cup [k.l]$. (Thus, $R = R_{[1,n]}$).

1. $R_{[i,j]\cup[k,l]}$ is a *TGB structure* (three-groups-of bands structure) if either

   (a) $[i,j] \cup [k,l]$ is a gapped region, $R_{[i,j]\cup[k,l]}$ contains base pair(s) $i.l$ and $j.k$, any base pair of $R_{[i,j]\cup[k,l]}$ which spans the gapped region $[i,j] \cup [k,l]$ does not cross any other base pair of $R_{[i,j]\cup[k,l]}$, and any nested substructures $R_{[i,j]\cup[k,l]}$, in weakly closed regions $[i,j]$ or $[k,l]$ are CCJ structures.

   (b) $R_{[i,j]\cup[k,l]}$ can be decomposed into a CCJ structure and a TGB structure in one of the following ways, for some d:

Range of d	*CCJ structure*	*TGB structure*
$i < d \le j$	$R_{[i,d-1]}$	$R_{[d,j]\cup[k,l]}$
$i \le d < j$	$R_{[d+1,j]}$	$R_{[i,d]\cup[k,l]}$
$k < d \le l$	$R_{[k,d-1]}$	$R_{[i,j]\cup[d,l]}$
$k \le d < l$	$R_{[d+1,l]}$	$R_{[i,j]\cup[k,d]}$

   (c) $R_{[i,j]\cup[k,l]}$ can be decomposed into a band and a TGB structure in one of the following ways, for some d and e:

Range of d	*Range of e*	*Band borders*		*TGB structure*
$i \le d < j$	$i < e \le j$	$i.j$	$d.e$	$R_{[d+1,e-1] \cup [k,l]}$
$i \le d < j$	$k < e \le l$	$i.l$	$d.e$	$R_{[d+1,j] \cup [k,e-1]}$
$i < d \le j$	$k \le e < l$	$d.e$	$j.k$	$R_{[i,d-1] \cup [e+1,l]}$
$k \le d < l$	$k < e \le l$	$k.l$	$d.e$	$R_{[i,j] \cup [d+1,e-1]}$

2. $R_{[i,l]}$ is a *CCJ pseudoknot* if it is the union of two TGB structures $R_{[i,j] \cup [d+1,k]}$ and $R_{[j+1,d] \cup [k+1,l]}$.

3. Finally, $R_{[i,l]}$ is a *CCJ structure* if
   (a) $R_{[i,l]}$ is empty, or
   (b) $i.l \in R_{[i,l]}$ and $R_{[i+1,l-1]}$ is a CCJ structure, or
   (c) for some $k \in [i, l-1]$, $R_{[i,l]} = R_{[i,k]} \cup R_{[k+1,l]}$ and both $R_{[i,k]}$ and $R_{[k+1,l]}$ are CCJ structures, or
   (d) $R_{[i,l]}$ is a CCJ pseudoknot.

3. RECURRENCES

Our algorithm finds the minimum free energy CCJ structure for a given input sequence S. We express energy values for various MFE substructure types as recurrences. A dynamic programming algorithm can compute these energies and store them in arrays, and a standard backtracking approach can then deduce R from the arrays. In what follows, we focus on providing the needed recurrences.

3.1. W and V

The starting point for recurrences that express the MFE of pseudoknot free structures are the following two terms. The first is $W(i,l)$, the MFE of all structures $R_{i,l}$ over region $[i,l]$. The second is $V(i,l)$, the

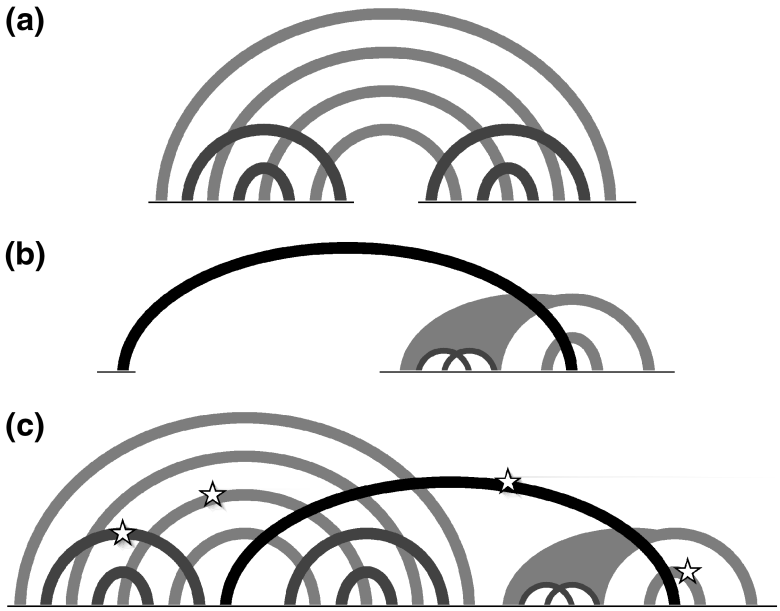

FIG. 2. TGB (three-groups-of-bands) and CCJ pseudoknotted secondary structures. (**a**) In this TGB structure for a gapped region, each shaded are represents a band—a set of pseudoknotted base pairs which span the arc. The left and right groups each have two bands, and the middle group has four bands. (The outermost middle "band" does not appear to be a band, since it does not cross any other bands, but as later parts of the figure show, once this TGB structure is overlaid with another TGB structure, all bands do indeed cross other bands.) (**b**) A structure with two groups of bands, and with pseudoknotted structures nested within one of these bands. In general, not all bands illustrated in part a need be present in a TGB structure, as long as at least one band is in the middle group (labeled with star in part c). (**c**) A CCJ structure is obtained by overlaying two TGB structures. This example is obtained by overlaying the structures of parts a and b. Embedded in this structure is a chain of four overlapping stems, which are labeled by four stars.

minimum free energy of all structures for region $[i,l]$ that contain base pair $i.l$. Here, we extend these to also include pseudoknotted structures. We set $W(i,l)=0$ if $i \geq l$. Otherwise,

$$W(i,l) = \min \left\{ \begin{array}{l} \min_{i \leq d < i} W(i,d-1) + V(d,l) \\ \min_{i \leq d < t} W(i,d-1) + P(d,l) + \gamma_0 \end{array} \right\}$$

$$V(i,l) = \min\{V_{hairpin}(i,l), V_{iloop}(i,l), V_{mloop}(i,l)\}$$

In the recurrence for $W(i,l)$, the first case arises when l is paired, say with base d, and $l.d$ is not pseudoknotted. Then the overall structure consists of two substructures, one in the region $[i,d-1]$ (handled by the W recurrence) and one in region $[d,l]$ which must contain base pair $d.l$ (handled by the V recurrence). The second case arises when l is the rightmost paired base in a pseudoknot (handled by the P recurrence). This is an external pseudoknot, therefore we add a γ_0 penalty. The recurrence for $V(i,l)$ minimizes over three terms, which handle hairpin loops, internal loops, and multiloops, respectively. Details for these types of loops are given in a later section; the term which handles multiloops takes into account the possibility that pseudoknots are nested in base pair $i.l$.

3.2. Pseudoknots

$P(i,l)$ is the minimum free energy of a CCJ pseudoknot for region $[i,l]$, not counting the initiation penalty. If $i \geq l$, $P(i,l) = +\infty$. Otherwise, the following recurrence expresses $P(i,l)$ using three intermediate points j, d, and k. These points, together with i and l, define two gapped regions, namely $[i,j] \cup [d+1,k]$ and $[j+1,d] \cup [k+1,l]$. $P(i,l)$ is the sum of contributions from two gapped regions:

$$P(i,l) = \min_{i \leq j < d < k < l} PK(i,j,d+1,k) + PK(j+1,d,k+1,l).$$

$PK(i,j,k,l)$ is the minimum free energy of a TGB structure $R_{[i,j] \cup [k,l]}$ for the gapped region $[i,j] \cup [k,l]$, given that both i and l are paired (not necessarily to each other) and the pairs involving i and l are not part of nested substructures, assuming also that some base pair (which is not in $R_{[i,j] \cup [k,l]}$) crosses the gap $[j+1,k-1]$.

The recurrence for PK uses terms P_L, P_M, P_O, and P_R. Informally, P_L and P_R handle bands in the left and right groups of the TGB structure, respectively. Both P_M and P_O are needed to handle bands in the middle group. More precisely, $P_L(i,j,k,l)$ is the minimum free energy of a TGB structure in gapped region $[i,j] \cup [k,l]$, excluding from this energy the term $\gamma_2(i,j)$ (which is accounted for in PK), given that i is paired with j, k is paired and the pair involving k is not in a nested substructure, and l is paired (not necessarily to k) and the pair involving l is not in a nested substructure. $P_R(i,j,k,l)$, $P_O(i,j,k,l)$, and $P_M(i,j,k,l)$ are defined similarly over gapped region $[i,j] \cup [k,l]$. For $P_R(i,j,k,l)$, k must pair with l, and i and j are paired (not necessarily to each other) and are not in nested substructures. For $P_O(i,j,k,l)$, i must pair with l, and j and k are paired (not necessarily to each other) and are not in nested substructures. Finally, for $P_M(i,j,k,l)$, j must pair with k, and i and l are paired (not necessarily to each other) and are not in nested substructures.

Note that $PK(i,j,k,l)$ only requires that i and l are paired, whereas P_L, P_M, P_O, and P_R require that all four indices are paired. The first two lines of the recurrence below for PK allow the indices j and k to be shifted to a position at which they are paired. The WP terms handle nested substructures in a pseudoloop. The remaining lines handle bands in the left, right, or middle groups of the MFE structure. These three lines have a γ_2 term, to account for the energy contribution of a border of a band (which is a closing pair of a pseudoloop).

If it is not the case that $i \leq j < k-1 < l$, then

$$PK(i,j,k,l) = P_L(i,j,k,l) = P_M(i,j,k,l) = P_R(i,j,k,l) = P_O(i,j,k,l) = +\infty.$$

Otherwise,

$$PK(i,j,k,l) = \min \left\{ \begin{array}{l} \min_{i < d < j} PK(i,d,k,l) + WP(j+1,d-1) \\ \min_{k < d < l} PK(i,j,d,l) + WP(d+1,k-1) \\ P_L(i,j,k,l) + \gamma_2(i,j) \\ P_M(i,j,k,l) + \gamma_2(k,j) \\ P_R(i,j,k,l) + \gamma_2(k,l) \\ P_O(i,j,k,l) + \gamma_2(i,l) \end{array} \right\}$$

$$P_L(i,j,k,l) = \min\left\{\begin{array}{l} P_{L,iloop}(i,j,k,l) \\ P_{L,mloop}(i,j,k,l) \\ P_{fromL}(i+1,j-1,d,l)+\gamma_2(j,i) \end{array}\right\}$$

$$P_R(i,j,k,l) = \min\left\{\begin{array}{l} P_{R,iloop}(i,j,k,l) \\ P_{R,mloop}(i,j,k,l) \\ P_{fromR}(i,j,k+1,l-1)+\gamma_2(l,k) \end{array}\right\}$$

$$P_M(i,j,k,l) = \min\left\{\begin{array}{l} P_{M,iloop}(i,j,k,l) \\ P_{M,mloop}(i,j,k,l) \\ P_{fromM}(i,j-1,k+1,l)+\gamma_2(j,k) \\ \quad \gamma_2(i,l), \text{if} i=j \text{ and } k=l \end{array}\right\}$$

$$P_O(i,j,k,l) = \min\left\{\begin{array}{l} P_{O,iloop}(i,j,k,l) \\ P_{O,mloop}(i,j,k,l) \\ P_{fromO}(i+1,j,k,l-1)+\gamma_2(l,i) \end{array}\right\}$$

The first two rows of the P_L recurrence handle the cases where $i.l$ is the outer closing base pair of an internal loop or multiloop that span a band. The third row handles the case where $i.l$ is the inner border of a band.

Figure 3 shows how these recurrences unwind to calculate the energy of a kissing hairpin structure. Figure 4 illustrates a more general case, which includes a chain of four bands.

3.3. Transitioning between band groups in pseudoknots

In the above recurrences for P_L, P_R, P_M, and P_O, terms P_{fromL}, etc., are used, to handle transitions from base pairs in one group to base pairs in another group. If transitioning from P_L via $P_{fromL}(i,j,k,l)$, then we need to allow for nested substructures either to the left of index j or to the right of index i, or both. The first two lines of the next recurrence allow for these possibilities, and ensure that when $P_R(i,j,k,l)$, $P_M(i,j,k,l)$, or $P_O(i,j,k,l)$ are called, i and j are at positions where they are paired (not necessarily with each other) and the pairs are not part of nested substructures.

$$P_{fromL}(i,j,k,l) = \min\left\{\begin{array}{l} \min_{i<d<j} P_{fromL}(d,j,k,l)+WP(i,d-1) \\ \min_{i<d<j} P_{fromL}(i,d,k,l)+WP(d+1,j) \\ P_R(i,j,k,l)+\gamma_2(k,l) \\ P_M(i,j,k,l)+\gamma_2(k,j) \\ P_O(i,j,k,l)+\gamma_2(i,l) \end{array}\right\}$$

$$P_{fromR}(i,j,k,l) = \min\left\{\begin{array}{l} \min_{k<d<l} P_{fromR}(i,j,d,l)+WP(k,d-1) \\ \min_{k<d<l} P_{fromR}(i,j,k,d)+WP(d+1,l) \\ P_M(i,j,k,l)+\gamma_2(k,j) \\ P_O(i,j,k,l)+\gamma_2(i,l) \end{array}\right\}$$

$$P_{fromM}(i,j,k,l) = \min\left\{\begin{array}{l} \min_{i<d<j} P_{fromM}(i,d,k,l)+WP(d+1,j) \\ \min_{k<d<l} P_{fromM}(i,j,d,l)+WP(k,d-1) \\ P_L(i,j,k,l)+\gamma_2(i,j) \\ P_R(i,j,k,l)+\gamma_2(k,l) \\ P_O(i,j,k,l)+\gamma_2(i,l) \end{array}\right\}$$

$$P_{fromO}(i,j,k,l) = \min\left\{\begin{array}{l} \min_{i<d<j} P_{fromO}(i,d,k,l)+WP(i,d-1) \\ \min_{k<d<l} P_{fromO}(i,j,d,l)+WP(d+1,l) \\ P_L(i,j,k,l)+\gamma_2(i,j) \\ P_R(i,j,k,l)+\gamma_2(k,l) \end{array}\right\}$$

Note that in P_{fromR}, it's not possible to transition to P_L. This is because the recurrences are designed so that bands are handled in rounds. Within a round, bands in the left are handled first, if any, then those in the right, if any, and then those in the middle, with bands handled by P_M (if any) handled before those handled by P_O. A middle band *must* be handled in each round; otherwise, for example, two ''bands'' in the left group, added in different rounds, would collapse into one, causing the recurrences to incorrectly add γ_2

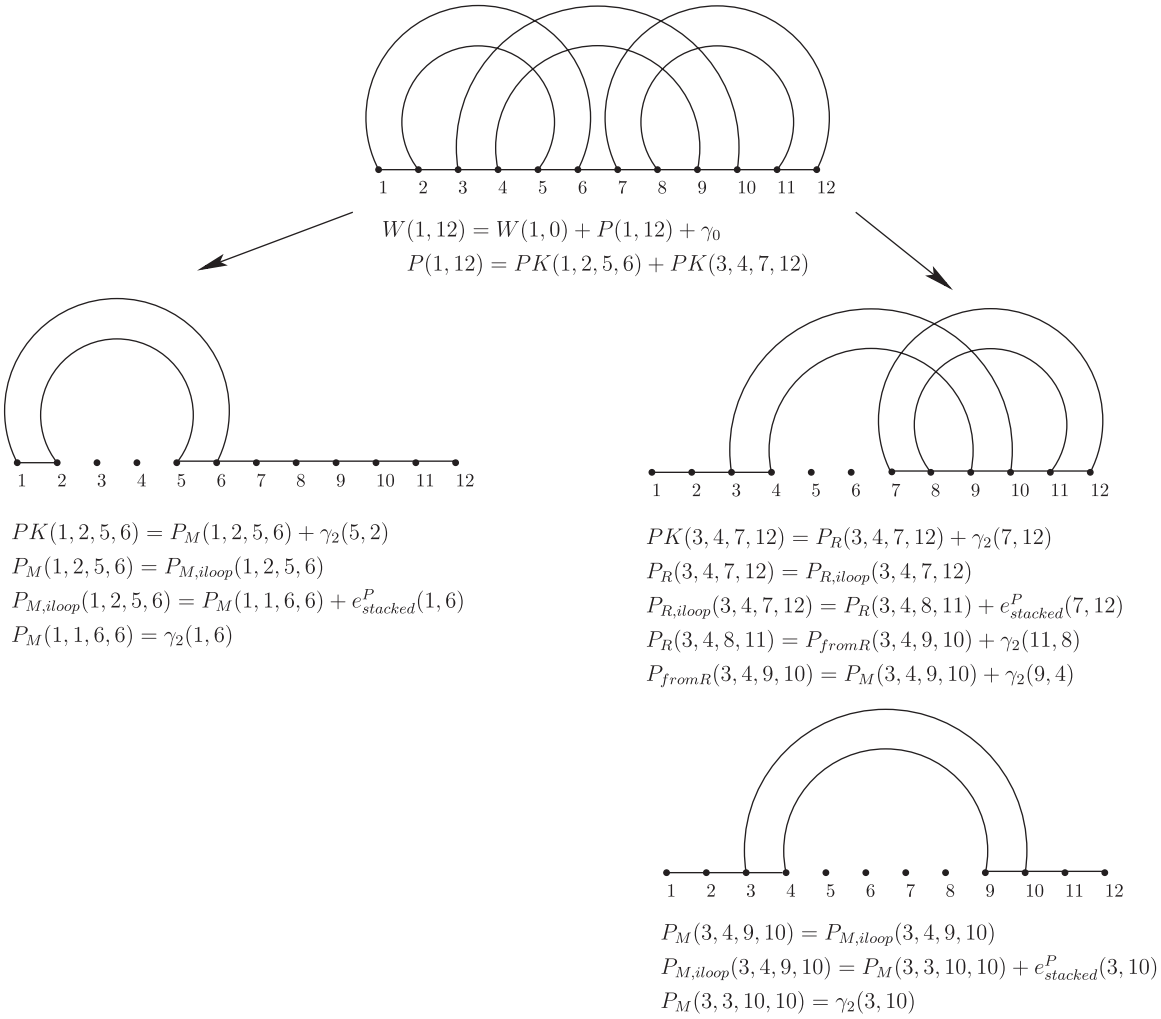

FIG. 3. Illustration of how the W recurrence unwinds, when the MFE structure for a sequence of length 12 is the simple kissing hairpin structure illustrated at the top of the figure. Since the overall structure is pseudoknotted, it is handled by the second case of the W recurrence. Here we have $W(1,0) = 0$; therefore, the energy is accounted for by $P(1,12) + \gamma_0$. In the P recurrence, the structure is divided into two TGB structures, namely $PK(1,2,5,6)$ and $PK(3,4,7,12)$. The term $PK(1,2,5,6)$ takes care of the internal loop of the gapped region $[1,2] \cup [5,6]$, and $PK(3,4,7,12)$ calculates the energy of the rest of the structure by transitioning between the P_R, $P_{R,iloop}$, P_{fromR}, P_M, and $P_{M,iloop}$ recurrences.

terms for band "borders" that are not actually borders. For this reason, no row in P_{fromR} has a P_L term, and so a band in the left group cannot be handled directly after a band in the right group. Also, P_{fromO} does not have a row with a P_M term, to ensure that P_M cannot be used twice in the same round.

We need a base case, when $i = j$ and $k = l$. In this case, we need to provide a way to exit the recurrences, with the last added base pair being $i.l(=j.k)$. If the recurrences are exited from P_M via P_{fromM}, then energy costs associated with the last band (which includes the base pair $i.l$) have already been accounted for. Therefore, $P_{fromM}(i,j,k,l) = 0$ when $i = j$ and $k = l$. Similarly, $P_{fromO}(i,j,k,l) = 0$. If the recurrences are exited from P_L or P_R (via P_{fromL} or P_{fromR}), we need to add energy costs for the base pair $i.l$, which forms a band of its own. We do this as follows: when $i = j$ and $k = l$,

$$P_{fromR}(i,j,k,l) = P_{fromL}(i,j,k,l) = \gamma_2(j,k) + \gamma_2(k,j).$$

3.4. Nested substructures

Substructures can be nested in different types of loops. We let $WM(i,l)$, $WB(i,l)$, and $WP(i,l)$ denote the minimum free energies of all structures $R_{[i,l]}$ for region $[i,l]$ that are nested in a multiloop that does not span

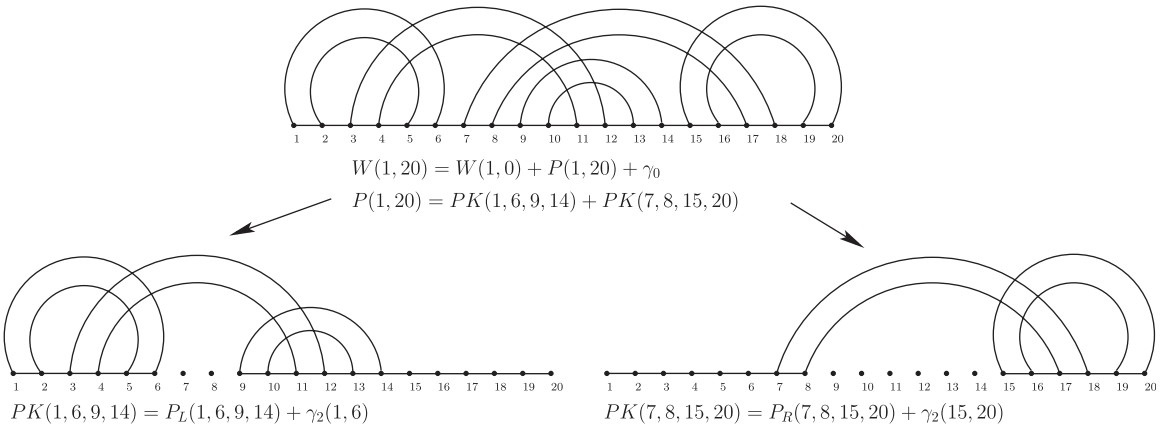

FIG. 4. Illustration of how the W recurrence unwinds, when the MFE structure for a sequence of length 20 is the structure illustrated at the top of the figure. Note that this structure has five interleaved bands. As in Figure 3, the overall structure is pseudoknotted; thus, it is handled by $P(1,20)+\gamma_0$. Here $P(1,20)$ is divided into $PK(1,6,9,14)$, and $PK(7,8,15,20)$. The TGB structure of $PK(1,6,9,14)$ is similar to Figure 3, and is handled similarly, with the only difference being that the leftmost band is handled by P_L recurrence instead of a PK recurrence. The TGB structure of $PK(7,8,15,20)$ is similar to the right part of Figure 3.

a band, a multiloop that spans a band, and a pseudoloop, respectively. $WM'(i,l)$ and WB' (i,l) are similar to $WM(i,l)$ and $WB(i,l)$, respectively, except that the region $[i,l]$ must contain at least one base pair. If $i>l$ then

$$WM(i,l) = WB(i,l) = WP(i,l) = 0, \text{ and }, WM'(i,l) = WB'(i,l) = +\infty$$

Otherwise,

$$WM(i,l) = \min\{WM'(i,l), \beta_1(l-i+1)\}$$

$$WM'(i,l) = \min\left\{\begin{array}{l}\min_{i\le d<e\le l} WM(i,d-1) + V(d,e) + \beta_1(l-e) + \beta_2(d,e) \\ \min_{i\le d<e\le l} WM(i,d-1) + P(d,e) + \beta_1(l-e) + \gamma_0^m\end{array}\right\}$$

$$WB(i,l) = \min\{WB'(i,l), \beta_1^p(l-i+1)\}$$

$$WB'(i,l) = \min\left\{\begin{array}{l}\min_{i\le d<e\le l} WB(i,d-1) + V(d,e) + \beta_1^p(l-e) + \beta_2^p(d,e) \\ \min_{i\le d<e\le l} WB(i,d-1) + P(d,e) + \beta_1^p(l-e) + \gamma_0^m\end{array}\right\}$$

$$WP(i,l) = \min\{WP'(i,l), \gamma_1(l-i+1)\}$$

$$WP'(i,l) = \min\left\{\begin{array}{l}\min_{i\le d<e\le l} WP(i,d-1) + V(d,e) + \gamma_1(l-e) + \gamma_2(d,e) \\ \min_{i\le d<e\le l} WP(i,d-1) + P(d,e) + \gamma_1(l-e) + \gamma_0^p\end{array}\right\}$$

In the above recurrences, the terms involving β_1, β_1^p, or γ_1 account for unpaired bases in ordinary multiloops, multiloops that span a band, and pseudoloops, respectively. The terms involving β_2 and β_2^p account for closing base pairs in multiloops, and the γ_0^m and γ_p initiation terms account for new pseudoknots that are nested in a multiloop or pseudoloop, respectively.

3.5 Internal loops and multiloops

First, we give the recurrences for hairpins, and for internal loops and multiloops that do not span a band. If $i>l-3$ then no such loop can be within base pair $i.l$, and so

$$V_{hairpin}(i,l) = V_{iloop}(i,l) = V_{mloop}(i,l) = +\infty$$

Otherwise,

$$V_{hairpin}(i,l) = e_H(i,l)$$

$$V_{iloop}(i,l) = \min_{i<d<e<l}(V(d,e) + e_{int}(i,d,e,l))$$

$$V_{mloop}(i,l) = \min_{i<d<e<l}\left\{\begin{array}{l} WM'(i+1,d-1) + V(d,e) + \beta_0 + \beta_1(l-e) + \beta_2(l,i) + \beta_2(d,e) \\ WM(i+1,d-1) + P(d,e) + \beta_0 + \beta_1(l-e) + \beta_2(l,i) + \gamma_0^m \end{array}\right\}$$

The first and second rows for V_{mloop} handle the first and second cases of multiloops described in Section 2, respectively.

Next are recurrences for internal loops and multiloops that span a band. Only the recurrences associated with the left group are listed. The recurrences for the middle and right groups are similar.

$P_{L,iloop}(i,j,k,l)$ is the minimum free energy of a TGB structure in gapped region $[i,j] \cup [k,l]$ (excluding the term $\gamma_2(i,j)$), given that $i.j$ is the closing base pair of an internal loop that spans a band. The first row of the recurrence handles the case that this internal loop is a stacked pair, and the second row handles all other types of internal loops. The α_0^p and $\alpha_2^p(j,i)$ terms account for the initiation and closing base pair penalties, and $P_{L,iloop5}(i,j,k,l)$ accounts for penalties associated with unpaired bases and asymmetry. $P_{L,iloop5}$ has five indices but during the execution of the algorithm, only a portion of the corresponding array, of size $O(n^4)$, is kept in memory at any given time, thus the space requirement is still $O(n^4)$ (see Section 3.6).

Similarly, $P_{L,mloop}(i,j,k,l)$ is the minimum free energy of a TGB structure in gapped region $[i,j] \cup [k,l]$ (excluding the term $\gamma_2(i,j)$), given that $i.j$ is the closing base pair of a multiloop that spans a band. Suppose d is the base with the smallest index, such that $i<d$ and $d.e$ spans the same band as $i.j$ for some e. Then, the first row of the recurrence for $P_{L,mloop}$ handles the case that there is a nested substructure in region $[i,d]$. The second case allows for the possibility that there is no nested substructure in region $[i,d]$, in which case there must be a nested substructure in region $[e,j]$.

$$P_{L,iloop}(i,j,k,l) = \min\left\{\begin{array}{c} P_L(i+1,j-1,k,l) + e^P_{stacked}(i,i+1,j-1,j) \\ \min_{0\le s\le j-i-7}(P_{L,iloop5}(i,j,k,l,s) + \alpha_0^p + \alpha_2^p(j,i)) \end{array}\right\}$$

$$P_{L,iloop5}(i,j,k,l,s) = \min\left\{\begin{array}{c} P_{L,iloop5}(i+1,j-1,k,l,s-2) + \alpha_1^p(s) - \alpha_1^p(s-2) \\ P_L(i+s+1,j-1,k,l) + \alpha_1^p(s) + \alpha_3^p(s) + \alpha_2^p(i+s1,j-1) \\ P_L(i+1,j-s-1,k,l) + \alpha_1^p(s) + \alpha_3^p(s) + \alpha_2^p(i+1,j-s-1) \end{array}\right\}$$

$$P_{L,mloop}(i,j,k,l) = \min_{i<d<j-1}\left\{\begin{array}{l} P_{L,mloop0}(d,j,k,l) + WB'(i+1,d-1) + \beta_0^p + \beta_2^p(j,i) \\ P_{L,mloop1}(d,j,k,l) + WB(i+1,d-1) + \beta_0^p + \beta_2^p(j,i) \end{array}\right\}$$

$$P_{L,mloop0}(i,d,k,l) = \min_{d<e<j}(P_L(d,e,k,l) + WB(e+1,j-1) + \beta_2^p(d,e))$$

$$P_{L,mloop1}(i,d,k,l) = \min_{d<e<j}(P_L(d,e,k,l) + WB'(e+1,j-1) + \beta_2^p(d,e))$$

3.6. Algorithm

A dynamic programming algorithm computes the solutions to the above recurrences in the following way. First, all energies which are base cases are calculated. Then, energies are computed according to the following schedule. For each value of T, $T = 2, 3, \ldots, n$, where n is the length of the input RNA sequence,

- compute all energies over regions $[i,j]$ with $j-i+1=T$;
- compute all energies over gapped regions $[i,j] \cup [k,l]$ with $j-i+l-k+2=T$;
- compute $P_{X,iloop5}(i,j,k,l,s)$ with $j-i+l-k+2=T$, for each X in $\{L,R,M,O\}$ and for all $s, 1 \le s \le n$;
- discard $P_{X,iloop5}(i,j,k,l,s)$ with $j-i+l-k+2=T-3$, for each X in $\{L,R,M,O\}$ and for all $s, 1 \le s \le n$.

The time requirement is $O(N^5)$ since all energies of the form $P_{X,iloop5}(i,j,k,l,s)$ can be computed in constant time and all other energies can be computed in at most linear time from energies computed earlier in the schedule. All energies computed are saved except for those discarded as described in the schedule, resulting in a space requirement of $\Theta(n^4)$.

4. CONCLUSION

In this article, we have provided a new algorithm for calculating the minimum free energy of pseudoknotted RNA secondary structures. Our algorithm runs in $O(n^5)$ time and $O(n^4)$ space, and handles a more general class of structures than previous algorithms with the same time and space bounds. A biologically important structure that can be handled by our algorithm, but not by previous $O(n^5)$ algorithms, is

the aptamer core of a SAM-IV riboswitch (Weinberg et al., 2008). Unlike previous $O(n^5)$ algorithms, our algorithm can also handle kissing hairpin structures with nested substructures, and also chains of four stems.

There are several interesting directions for future work. First, we note that Reeder and Giegerich (2004) have shown that by constraining some aspects of secondary structures, such as the lengths of some pseudoknotted stems, it is possible to recognize H-type pseudoknots in $\Theta(n^4)$ time and $\Theta(n^2)$ space. Could the space or time of our algorithm be reduced, using methods similar to those of Reeder and Giegerich, allowing more restricted types of kissing hairpin structures to be handled more efficiently? A preliminary analysis of structures in the RNA STRAND database (Andronescu et al., 2008) indicates that the restrictions imposed by Reeder and Giegerich's methods would not exclude many biologically important structures, so their approach is well motivated from a practical standpoint.

In a similar vein, we also note that our algorithm could be generalized further while still keeping the time at $\Theta(n^5)$ and space at $\Theta(n^4)$. For example, bands consisting of one isolated base pair or a short stem (of constant length, independent of the length of the input sequence) could be handled without minimizing over a linear-sized range, allowing for chains of short stems of arbitrarily large length. However, we conjecture that the class of algorithm of Rivas and Eddy (1999) is strictly more general than any class that can be handled in $O(n^5)$ time and $O(n^4)$ space.

We expect that the algorithm presented in this article can be adapted to calculate a partition function for pseudoknotted structures that is more general than the partition function of Dirks and Pierce (2003). One of the authors of this paper (Chen) plans to design and implement such an algorithm.

ACKNOWLEDGMENTS

We thank Mirela Andronescu for her input on features of pseudoknotted structures in the RNA STRAND database and Takachi Yokomori for explanations of the structures that can be handled by the algorithm of Uemura et al (1999).

DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Akutsu, T. 2000. Dynamic programming algorithms for RNA secondary structure prediction with pseudoknots. *Discrete Appl. Math.* 104, 45–62.

Andronescu, M., Bereg, V., Hoos, H., et al. 2008. RNA STRAND: the RNA secondary structure and statistical analysis database. *BMC Bioinform*, 9, 340.

Dirks, R.M., and Pierce, N.A. 2003. A partition function algorithm for nucleic acid secondary structure including pseudoknots. *J. Comput. Chem.* 24, 1664–1677.

Dirks, R.M., and Pierce, N.A. 2004. Triggered amplification by hybridization chain reaction. *Proc. Natl. Acad. Sci. USA* 101, 15275–15278.

He, Y., Ye, T., Su, M., et al. 2008. Hierarchical self-assembly of DNA into symmetric supramolecular polyhedra. *Nature* 452, 198–201.

Hofacker, I.L., Fontana, W., Stadler, P.F., et al. 1994. Fast folding and comparison of RNA secondary structures. *Monatshefte für Chemie/Chemical Monthly* 125, 167–188.

Lee, K., Varma, S., Santalucia, J., et al. 1997. *In vivo* determination of RNA structure-function relationships: analysis of the 790 loop in ribosomal RNA. *J. Mol. Biol.* 269, 732–743.

Lyngsø, R.B. 2004. Complexity of pseudoknot prediction in simple models. In Diaz, J., Karhumäki, J., Lepistö, A., and Sannella, D., eds. ICALP, Springer, *Lect. Notes Comput. Sci.* 3142, 919–931.

Lyngsø, R.B., and Pedersen, C.N. 2000. RNA pseudoknot prediction in energy-based models. *J. Comput. Biol.* 7, 409–427.

Mathews, D.H., Sabina, J., Zuker, M., et al. 1999. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.* 288, 911–940.

McCaskill, J.S. 1990. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers* 29, 1105–1119.

Rastegari, B., and Condon, A. 2007. Parsing nucleic acid pseudoknotted secondary structure: algorithm and applications. *J. Comput. Biol.* 14, 16–32.

Reeder, J., and Giegerich, R. 2004. Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynamics. *BMC Bioinform.* 5.

Rivas, E., and Eddy, S.R. 1999. A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J. Mol. Biol.* 285, 2053–2068.

Rothemund, P.W., Papadakis, N., and Winfree, E. 2004. Algorithmic self-assembly of DNA Sierpinski triangles. *PLoS Biol.* 2, e431.

Simmel, F.C., and Dittmer, W.U. 2005. DNA nanodevices. *Small* 1, 284–299.

Uemura, Y., Hasegawa, A., Kobayashi, S., et al. 1999. Tree adjoining grammars for RNA structure prediction. *Theor. Comput. Sci.* 210, 277–303.

Weinberg, Z., Regulski, E.E., Hammond, M.C., et al. 2008. The aptamer core of SAM-IV riboswitches mimics the ligand-binding site of SAM-I riboswitches. *RNA* 14, 822–828.

Zuker, M. 2003. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* 31, 3406–3415.

Zuker, M., and Sankoff, D. 1984. RNA secondary structures and their prediction. *B Math Biol.* 46, 591–621.

Zuker, M., and Stiegler, P. 1981. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.* 9, 133–148.

Address reprint requests to:
Hosna Jabbari
Department of Computer Science
University of British Columbia
201-2366 Main Mall
Vancouver, BC, V6T 1Z4, Canada

E-mail: hjabbari@cs.ubc.ca